Uncertainty-Aware Fuzzy Neural Networks for Explainable Risk Scoring in Real-Time Payment Transactions

Hasan Huseynov[1] and Prof. Dr. Shahnaz Shahbazova[2] and Assoc. Prof. Dr. Amir Mosavi[3]

[1] Azerbaijan State University of Economics, Baku, Azerbaijan
hesenhuseynov372@gmail.com
[2] Azerbaijan State University of Economics, Baku, Azerbaijan
shahbazova@gmail.com
[3] Obuda University, Budapest, Hungary
amir.mosavi@uni-obuda.hu

Abstract. UA-FNN integrates an ANFIS-inspired fuzzification and rule-firing pipeline with a two-layer LSTM consequent network to extract temporal patterns, Monte Carlo Dropout to estimate epistemic uncertainty, and a new Fuzzy-SHAP post-processor which transforms SHAP DeepExplainer attributions into Gaussian membership-function-weighted linguistic tokens. The results of each inference pass are a calibrated risk score, a predictive uncertainty band sigma, a routing action, and a list of languages risk factors in order of rank meeting the requirements of an adverse-action explanation. On three benchmark data sets, the IEEE-CIS Fraud Detection dataset (590,540 transactions), ULB Credit Card Fraud (284,807 transactions), and synthetic PaySim (6.36 million transactions), UA-FNN with AUC-ROC of 0.971, 0.983, and 0.977 and F1-scores The Expected Calibration Error (ECE = 0.031) is 65% lower than the MC-Dropout DNN baseline. The need of the LSTM consequent and Gaussian membership functions is proved in a systematic ablation study of eight architectural variants. Experiments on uncertainty routing demonstrate that transactions characterized by high uncertainty with 17.5% of total volume have a rate of fraud 20-32 times greater than transactions with low uncertainty, making predictive uncertainty a first-class risk management predictive. All findings are confirmed using DeLong pairwise AUC and Wilcoxon signed-rank tests with $p < 0.01$.

Keywords: uncertainty-aware machine learning; fuzzy neural networks; ANFIS; Monte Carlo Dropout; SHAP

1 Introduction

The globe payment platform of digital payments is now processing over 3.4 billion card transactions every year with total amount processing over USD 1.8 trillion [1]. The number of the fraud has grown, US Federal Trade Commission has registered at least 425,977 records of credit card fraud in the year 2023 alone, and the volume of money stolen via cards is more than USD 32 billion in the year 2023 [1, 16]. There has never been such pressure on payment processors, issuing banks and payment service providers to deploy proper, low-latency and transparently responsible fraud detection systems as they are today. Current-day models of production fraud detection primarily come in either the form of gradient-boosted tree ensembles or neural networks. These models have high AUC-ROC scores that have three endemic weaknesses. They are opaque: the rationale behind the decisions cannot be expressed in such a manner that may be grasped by compliance officers, customers, and regulators. They produce uncalibrated scores of confidence: a model may produce a decision which is probable, but far, in its training distribution, and that is just what occurs when the rip-offs construct new adversarial patterns [3]. The automated payment scoring and credit scoring fall under high-risk AI systems in the EU AI Act (Regulation 2024/1689), which will take effect in August 2024, in Annex III. It necessitates elucidation, human regulation, documenting precision, and threat regulation procedures [20]. Article 86 codifies the right of an individual to be informed of any decision made by AI with a material impact on an individual. The GDPR article 22 also sets the same requirements on automated individual decision-making. These specifications have to be followed by requiring fraud detection systems which are precise, comprehensible, and well-calibrated simultaneously. Adaptive Neuro-Fuzzy Inference System (ANFIS) proposed by Jang (1993) [4] took the approximation properties of the neural networks and interpolated the explicit rule format of the Sugeno-type fuzzy inferences systems. ANFIS has been found to demonstrate superb performances in the domain of financial default forecasting [15] and bank risk rating [14]. The classical ANFIS architectures are however formulated to handle low-dimensional and stationary regression problems and cannot handle high dimensional non-stationary streaming transaction data, distributional shift calibrated epistemic uncertainty or even post-hoc feature-level explanations required in production deployment.

Research gaps. A literature review (Section 2 and Table 1) indicates that none of the existing literature fulfil all five criteria of a regulation-ready real-time payment risk engine: high predictive performance under class imbalance (R1); calibrated probabilistic uncertainty with epistemic decomposition (R2); linguistically interpretable, rule-based explanations that meet the GDPR and EU AI Act requirements (R3); fuzzy reasoning to capture transaction.

2 Problem Statement

Although much progress in machine learning-based fraud detection has already been achieved, none of the existing systems meets all five conditions of a regulation-ready, real-time payment risk engine. It is not the lack of the effective models because gradient-boosted ensembles and deep neural networks already have the competitive AUC-ROC scores, but the lack of structural compatibility between the need of modern fraud detection and the present architectures that can effectively provide this.

The problem is characterized in three aspects of incompatibility. To begin with, when it comes to model confidence that is uncalibrated, no accuracy suffices. A scoring system which assigns a transaction a high-level risk but does not measure the uncertainty of the score gives no information to support intelligent routing, human escalation, or a risk-based decision. The current fraudsters are continuously changing behavioral patterns, moving transactions to areas of the feature space not well represented during training, but current models do not distinguish between predictions with high and low confidence based on the distributions. Second, the EU AI Act and GDPR Article 22 put a legal requirement to describe individual automated decisions in a human-readable form. The raw neural network feature attributions are not compliant adverse-action explanations; they are dependent on an interjective translation layer that is not provided by existing fraud detection designs. Third, classical interpretable architectures such as ANFIS, although providing clear structures of rules, are limited to low-dimensional held inputs and does not model temporal behavioral patterns, high-dimensional available spaces, and concept drift as a feature of production payment streams.

The result is a trilemma: accurate systems cannot be interpreted; systems that can be interpreted cannot deal with complexity over time; and systems that measure uncertainty cannot run on-the-fly. Closing this gap and building a unified end-to-end architecture that is both discriminative, uncertain, linguistically explainable, and temporally expressive, and capable of inferences in sub-10 milliseconds is the key issue that this work is trying to accomplish. The Uncertainty-Aware Fuzzy Neural Network (UA-FNN) suggested to address this trilemma is created with this purpose.

3 Related Work

3.1 Sequential and Graph-Based Fraud Detection

The concept of payment fraud detection has evolved the rule-based threshold systems into state-of-the-art machine learning pipelines over the past 20 years. Realistic class-imbalance modelling and temporal train-test splits are the main factors that prove useful in this context because Dal Pozzolo et al. (2017) [5] have shown that naïvely over-sampled assessments largely overestimate their reported performance. Jurgovsky et al. (2018) [6] demonstrated that sequence models based on LSTMs and sensitive to the temporal context of card histories work better than the feature-engineering-intensive tabular ones on recall-oriented measures. The ensemble techniques [7] with attention weights have shown that the variance reduction of component models would be an

approximation of uncertainty estimation in a non-committal fully probabilistic model. Duan et al.

3.2 Uncertainty Quantification in Financial Machine Learning

The first systematic comparison of MC-Dropout, deep ensembles, and ensemble MC-Dropout on the ULB credit card fraud dataset was done by Habibpour et al. (2023) [3]. Their work proves that ensemble-based UQ performs better than single-model dropout on both calibration error and selective prediction accuracy under abstention, which forms a significant baseline to this study. Formalising MC-Dropout as approximate Bayesian inference in deep Gaussian processes, Gal and Ghahramani (2016) [22] gave the theoretical basis to the uncertainty component of UA-FNN. As demonstrated by Lakshminarayanan et al. (2017) [23], deep ensembles have superior calibration at the cost of inference costs which are linear in the size of the ensemble, rendering them incompatible with real-time card authorisation latency constraints.

3.3 Explainable AI for Payment Risk and Credit Scoring

SHAP [10] and LIME [16] are the two most common explainability paradigms of financial ML. Arrieta et al. (2021) [25] in a systematic literature review that identifies 150 studies on XAI in finance indicate that SHAP has axiomatic Shapley-value guarantees and LIME has the lowest speed. Hjelkrem and de Lange (2023) [11], in a study of a SHAP explanation on top of an open banking deep-learning credit scoring model on open banking textual data, have discovered that practitioner-generated explanation and expert judgement concurred, but post-processing (specialized domain information) was needed to generate regulation adverse-action notices. It was already known that SHAP already is the largest explanation tool used throughout the UK financial institutions as an analysis performed by a Financial Stabilizer Board (2024). Raw neural network feature SHAP attributions can however not be either directly interpreted by a non-technical human nor adversarial action notice without an intermediate mapping layer, exactly the missing link which the Fuzzy-SHAP bridge as suggested in the following fills.

3.4 Fuzzy Inference Systems in Financial Risk

The history of Fuzzy logic in financial risk management is long. Ahmed et al. (2023) [14] used ANFIS with membership functions that are Gaussian to a bank risk index Mahalanobis-distance based, and obtained sensitivity indices whose ranking of the predictive significance of a single financial ratio in a directly interpretable format. In their study of ANFIS and other algorithms, including logistic regression, ANN and FIS as predictors of Islamic bank credit default, Ayed and Bougatef (2024) [15] determined that ANFIS performed the best in prediction of credit default in the form of AUC (0.90) and misclassification cost (lowest), making it clear that the interpretable rule structure did not degrade predictive. (2024) [2] tested a time-series and neural network model that can take the concept drift into account in patterns of payment frauds, but without measuring uncertainty, and without explicit explainability. None of the literature combines ANFIS and deep temporal feature extraction with MC-Dropout UQ and SHAP-based linguistic explanation in a real-time compatible structure. Table 1 places UA-FNN in perspective to its nearest related close art.

Table 1. Systematic comparison of related work. '–' denotes absent capability; 'Partial' denotes partial support. CC = credit card; Pay = online payments. Bold row (proposed approach) satisfies all five requirements simultaneously.

Method / Model	XAI	UQ	Fuzzy	Real-Time	Dataset Domain
Gradient Boosting	–	–	–	Partial	CC Fraud
LSTM Sequence	–	–	–	Yes	CC Fraud
Ensemble + Attention	Partial	–	–	–	CC Fraud
Anomaly + Rule-Based	Partial	–	–	Yes	Online Pay
DNN + MC-Dropout	–	Yes	–	–	CC Fraud
CaT-GNN	Partial	–	–	Yes	CC Fraud
Time-Series NN + Fuzzy	–	–	Partial	–	CC Fraud
ANFIS + Mahalanobis	Yes	–	Yes	–	Banking Risk
ANFIS vs LR/ANN/FIS	–	–	Yes	–	Credit Default
Risk-Aware GNN	Partial	–	–	Yes	CC Fraud
Fuzzy-NN + LSTM + MC-Dropout + SHAP (Proposed)	**Yes**	**Yes**	**Yes**	**Yes**	**CC Fraud / Pay**

4 Proposed Methodology: UA-FNN

4.1 Problem Formulation

Every payment transaction reaching authorisation time t is represented by a feature vector xn which groups together d various attributes: The amount of money in the transaction on a log scale, the type of merchant, a distance travelled to date since the last purchase, the velocity of card usage over 1-hour, 6-hour and 24-hour periods, device fingerprint embeddings, an email domain risk score, and a set of time-based contextual indic The binary variable yn states that transaction is fraudulent or genuine (0 or 1). The aim is to fit a predictor f which takes this feature list to a fraud probability score r $_+$ as shown in Equation (1).

$$f : \mathbb{R}d \rightarrow [0,1] \text{ such that } \bar{r}_t = f(x_t) \approx P(y_t = 1 \mid x_t) \quad (1)$$

Compared to traditional models, which only generate a single risk score, UA-FNN is also capable of generating two other outputs in every transaction. The first one is the predictive uncertainty value σ n, a standard deviation of several forward passes with stochasticity under MC-Dropout, a reflection of the confidence the model has in its prediction. E$\mathscr{R}$ is a linguistic interpretation of SHAP that converts raw SHAP attribution values into understandable fuzzy terms a human can interpret. Lastly, the model is constructed based on realistic assumptions of how fraud behaviour changes with time: since the patterns underlining fraudulent transactions are changing at a constant rate, the system is modelled to change at periodic intervals, instead of viewing the problem as a static constant distribution.

4.2 Architecture Overview

UA-FNN is structured as a six-layer differentiable network. Layers L1 through L3 implement ANFIS-style fuzzification and rule-firing following the Takagi-Sugeno-Kang (TSK) fuzzy inference framework. Layer L4 replaces the classical linear TSK consequent with a two-layer LSTM sub-network. Layer L5 implements defuzzification with MC-Dropout stochastic inference. Layer L6 is a non-differentiable post-processor applying the Fuzzy-SHAP bridge.

4.3 Fuzzification Layer (L1)

The fuzzification layer transforms every raw numerical input feature to a sequence of soft membership levels stating the extent to which the observed value is a member of a specific linguistic category. Each input feature has five Gaussian membership functions in the domain of values of the feature, one that represents each of the linguistic levels: very low, low, medium, high, and very high. The reason why Gaussian functions are used is due to the smoothness and continuous derivability of the functions; thus, their parameters can be updated using backpropagation without the discontinuities in gradients brought about by piecewise functions like trapezoid.

Every Gaussian membership function has two learnable scalar parameters, a centre, defining the location in the feature domain around which the membership function is highest, and a width, which defines the extent to which the membership function is diffused around the centre. Instead of randomly initialising these parameters, a subtractive clustering algorithm is applied to the training data to discover natural clusters in the feature space, and the centroid and spread estimates of these clusters are used as starting values of each membership function. This warm start is based on data and converges faster and gives the first linguistic partitions to reflect the distribution of transaction features rather than the actual distribution of the transaction features before end-to-end fine-tuning.

Since there are five membership functions specified on each feature, the number of possible combinations of fuzzy rules is exponentially growing in the number of input dimensions and rapidly becomes computationally infeasible. This is addressed by a hierarchical sub-clustering step which ranks all candidate rules by a cluster-density product, which quantifies the denseness of the antecedent region of a rule in the training data, and only the top 128 rules are retained. This decreases the rule space of the

exponential number to a fixed and manageable active set without losing the rules best representing observed patterns of transactions.

4.4 Rule Firing and Normalization (L2-L3)

The rule firing layer determines how strongly each of the 128 active fuzzy rules is activated by a given transaction. Each rule specifies a particular combination of linguistic levels across all input features, and its firing strength is computed as the product of the individual membership degrees corresponding to those linguistic levels. Using a product aggregation operator means a rule fires strongly only when all of its linguistic conditions are simultaneously satisfied, capturing the conjunctive and semantics of a fuzzy rule antecedent.

Because raw firing strengths do not sum to one across rules, a normalisation step divides each rule's firing strength by the total sum of all firing strengths. After normalisation, the values sum exactly to one and behave like a probability distribution over the rule base, expressing each rule's relative contribution to the final output. This interpretable representation, where a normalised firing strength of 0.35 means a rule accounts for 35 percent of the model's reasoning on that transaction, is later exploited by the Fuzzy-SHAP bridge to construct regulation-compliant linguistic explanations.

4.5 LSTM Consequent Network (L4)

Classical ANFIS allocates a simple linear consequent to each fuzzy rule, unable to reflect the time dynamics of sequential transaction behaviour. UA-FNN substitutes this with a single two-layer LSTM network which takes a sliding window of the eight latest received card transactions and gives an output of a 64-dimensional hidden state which encapsulates the more recent behavioural context of the card such as velocity bursts, geographic drift, and changes in spending pattern.

Instead of training 128 distinct LSTMs, 1 set of LSTM parameters is used across all 128 rules. Every rule then implements a lightweight linear projection layer on top of the shared hidden state to give a rule-specific consequent scalar. This design is such that the overall number of parameters is reduced by the number of rules and the predictive nature of each rule is maintained. The joint temporal encoding and rule-specific projection is what leads to the UA-FNN discriminative advantage over non-discriminative (non-dynamic) tabular models, which ablation study in Section 5.2 validated.

4.6 Defuzzification and MC-Dropout Uncertainty Estimation (L5)

The defuzzification process is used to combine the 128 rule consequents into one scalar risk measure by summing the weighted contributions, with each consequent contribution being weighted by the normalised firing strength of the rule. Since the weights add to one, the outcome is a convex mixture confined within the region of the resultant values and attracted to the consequences of the rules that are firing strongly, which gives a continuous risk score that is directly interpretable as the estimated probability of fraud.

In order to measure certainty of prediction, the Dropout layers with dropout rate of 0.20 are used in LSTM and not turned off but are active at the inference time, the normal way. When 50 stochastic forward passes are run per transaction, the distribution of risk scores would be slightly different, as there would be random masking of a set of hidden

units each pass. The average of these 50 samples is the reported risk score and the standard deviation of these samples is the predictive uncertainty, which tells us how sensitive the output of the model is to small internal perturbations. When the standard deviation is large, it means that the transaction is close to a decision boundary or in an area it is not well represented during training.

4.7 The Fuzzy-SHAP Bridge (L6)

SHAP DeepExplainer initially estimates a feature attribution score of each input feature, which is the marginal contribution of the feature to a change in the expected risk score of the model, when the feature is altered to the average output of the model. The bridge overcomes this by computing, based on each input feature, the degree of highest activation of the five Gaussian membership functions of the observed feature value in that transaction. Such dominant membership role implies one of five linguistic levels and gives a natural language description of the manner in which that feature was acted. The name in the feature, level in a language, direction of attribution and the magnitude of the attribute are then amalgamated into a structured linguistic token. The three features that have the highest positive attributions are ranked and compiled into the final explanation, which is a concise and regulation-compliant adverse-action notice requiring no additional post-processing to be delivered to a customer or compliance record.

5 Experimental Setup

5.1 Datasets

Experiments were conducted across three publicly available benchmark datasets to assess generalisation across different data regimes.

- IEEE-CIS Fraud Detection: 590,540 transactions (Kaggle, Vesta Corp., 2019) with 3.5% fraud prevalence and 433 features including anonymised identity attributes, transaction properties, and derived velocity features. Primary evaluation dataset.
- ULB Credit Card Fraud: 284,807 European cardholder transactions (September 2013) with 0.172% fraud prevalence. Features are 28 PCA-transformed components, Time, and Amount. Tests performance under extreme class imbalance.
- PaySim (Synthetic Financial Transactions): 6,362,620 transactions generated by a multi-agent simulation calibrated to mobile money patterns (Lopez-Rojas, 2016). Fraud prevalence is 1.29%. Tests scalability and streaming latency.

A temporal split was applied for all datasets: the first 75% of the transaction time range was used for training and the remaining 25% for testing, respecting temporal ordering and preventing look-ahead bias following the methodology of Dal Pozzolo et al. [5].

5.2 Pre-processing and Feature Engineering

The pipeline that was used in all datasets was as follows. Numerical features were made standardised according to the training split where only the training data computed the statistics. Target-encoded categorical features were based on a 5-fold cross-validated smoothed mean estimate. Computation of velocity features was done in a streaming

consistent fashion and only past transactions were used. Missing goes have been filled with median imputation, mode imputation, and missingness indicators have been included as binary attributes. The problem of class imbalance was resolved using SMOTE oversampling [18] with a 1:10 fraction of fraud to legitimate ratio on the training sample, and inverse-frequency cost-weighted loss weighting.

5.3 Baselines

The pipeline that was used in all datasets was as follows. Numerical features were made standardised according to the training split where only the training data computed the statistics. Target-encoded categorical features were based on a 5-fold cross-validated smoothed mean estimate. Computation of velocity features was done in a streaming consistent fashion and only past transactions were used. Missing goes have been filled with median imputation, mode imputation, and missingness indicators have been included as binary attributes. The problem of class imbalance was resolved using SMOTE oversampling [18] with a 1:10 fraction of fraud to legitimate ratio on the training sample, and inverse-frequency cost-weighted loss weighting.

5.4 Evaluation Metrics and Statistical Testing

APUC-ROC, F1-score, Precision, Recall and False Positive Rate (FPR) are utilized as primary measures at the Youden J threshold. Calibration is measured through Expected Calibration Error (ECE) with M = 10 equal-width bins. The latency of inference is quantified at the 99 percentile with 10,000 concurrent transactions on a single-core CPU node (Intel Xeon E5-2690 v4, 2.60 GHz, 32 GB RAM, no GPU) indicating CPU-bound production. The difference between AUC-ROCs is tested by the DeLong test of the correlated ROC curves; the difference in F1, Precision, Recall and FPR is tested using the Wilcoxon signed-rank test over the 10 operation-time cross-validation. All p-values are two-tailed with significance levels: $p < 0.001$, $p < 0.01$, $p < 0.05$, ns = not significant.

6 Results and Discussion

6.1 Classification Performance

UA-FNN achieves the highest AUC-ROC (0.971), F1-score (0.921), Precision (0.934), and Recall (0.908), with an FPR of 1.22% and an ECE of 0.031. The AUC improvement over XGBoost (+0.029) is significant by DeLong test ($z = 4.81$, $p < 0.001$), and the improvement over MC-Dropout DNN (+0.013) is likewise significant ($z = 2.74$, $p < 0.01$). The 43% relative reduction in FPR relative to XGBoost translates to approximately 1,370 fewer false declines per 100,000 legitimate transactions, with direct customer experience and revenue implications. The Deep Ensemble baseline achieves competitive AUC (0.963) and ECE (0.048) but incurs 42.1 ms at the 99th percentile, exceeding real-time card authorization latency budgets. UA-FNN at 6.3 ms is well within the sub-10 ms target.

6.2 Ablation Study

Table 2 reports an eight-variant ablation study decomposing the contribution of each architectural component. The most consequential finding is that removing the LSTM consequent and reverting to a linear ANFIS consequent causes the largest performance drop ($\Delta AUC = -0.030$, $p < 0.001$), confirming that temporal transaction sequence modelling is the primary performance driver beyond the fuzzy structure. Replacing Gaussian MFs with Trapezoidal MFs causes a significant AUC reduction of 0.013 ($p < 0.01$), consistent with the theoretical advantage of differentiable Gaussian MFs for gradient-based optimisation. Halving the temporal window from $T = 8$ to $T = 4$ causes a significant AUC reduction of 0.011 ($p < 0.01$), while doubling to $T = 16$ provides no further improvement at substantially higher latency (9.8 ms), establishing $T = 8$ as the optimal context length. Removing MC-Dropout has a negligible effect on AUC (p = ns) but increases ECE by 119%, confirming that MC-Dropout provides calibration improvement rather than discriminative improvement.

Table 2. Ablation study on IEEE-CIS test set. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant. Δ AUC = signed AUC difference relative to the full UA-FNN model.

Variant	AUC-ROC	F1	FPR%	ECE	Lat.(ms)	Δ AUC vs Full
UA-FNN (Full Model)	**0.971**	**0.921**	**1.22**	**0.031**	**6.3**	**ref.**
w/o LSTM (linear consequent)	0.941	0.877	2.11	0.044	3.1	−0.030***
w/o MC-Dropout (point estimate)	0.969	0.918	1.28	0.068	4.9	−0.002 (ns)
w/o Fuzzy-SHAP (no explanation)	0.971	0.921	1.22	0.031	6.0	0.000 (ns)
Trapezoidal MFs (replace Gaussian)	0.958	0.901	1.51	0.039	6.4	−0.013**
Window T = 4 (half temporal context)	0.960	0.897	1.68	0.037	5.1	−0.011**
Window T = 16 (double context)	0.972	0.922	1.21	0.030	9.8	+0.001 (ns)
M = 64 rules (half rule base)	0.963	0.909	1.40	0.035	4.6	−0.008*
S = 20 MC samples	0.971	0.920	1.23	0.038	4.2	0.000 (ns)

6.3 Cross-Dataset Generalization

Table 3 reports AUC-ROC and F1-score across all three benchmarks. UA-FNN achieves consistent state-of-the-art performance: AUC-ROC of 0.983 on ULB (extreme imbalance, 0.17% fraud prevalence) and 0.977 on PaySim (large-scale synthetic), with DeLong-significant improvements over all baselines ($p < 0.01$). The generalisation of the LSTM consequent across the temporal structures of three fundamentally different datasets suggests that the hierarchical temporal-fuzzy feature extraction is robust to domain shift.

Table 3. Cross-dataset generalization results. DeLong-significant improvement over all baselines, $p < 0.01$. ULB = University of Liège dataset; PaySim = synthetic mobile money dataset.

Model	ULB AUC	ULB F1	IEEE-CIS AUC	IEEE-CIS F1	PaySim AUC	PaySim F1
XGBoost	0.969	0.847	0.942	0.871	0.954	0.862
LSTM (standard)	0.975	0.861	0.951	0.884	0.961	0.879
MC-Dropout DNN	0.978	0.874	0.958	0.893	0.966	0.891
UA-FNN (Proposed)	**0.983**	**0.912**	**0.971**	**0.921**	**0.977**	**0.908**

6.4 Qualitative Explainability Evaluation

Fuzzy-SHAP post-processor was tested in an organized test involving seven qualified fraud analysts (average experience 9.2 years, range 5 to 17 years), recruited in two European issuing banks. All the analysts were provided with 50 randomly selected fraudulent transaction explanations and rated them in terms of Accuracy, Actionability, and Clarity using a 5-point Likert-scale. Mean ratings were: Accuracy 4.31 (SD 0.67), Actionability 4.18 (SD 0.74), and Clarity 4.52 (SD 0.52). Analysts rated 94% of the explanations as accurate or very accurate and reported that the explicit uncertainty band in the explanation was helpful in the triage decision making process, with 86% saying that they would give high uncertainty cases first priority to review.

7 Limitations and Future Work

7.1 Architectural Limitations

UA-FNN operates on a predetermined length transaction window ($T = 8$), potentially missing ultra-long-range signals of fraud weeks or months, like account takeover preparation with a steady normalisation of behaviour. A natural extension is variable-length attention-based temporal aggregation based on a Transformer encoder. The parameterization of the Gaussian membership function is based on unimodal per-feature marginal distributions in each fuzzy partition; bimodal distributions due to heterogeneous sets of customers can be Type-2 fuzzy sets or adaptive hierarchical partitioning. Even though MC-Dropout approximates Bayesian inference that is principled, it is known to underestimate epistemic uncertainty when training data are dense [22], future research should compare MC-Dropout with normalizing flows and spectral-normalized Gaussian Processes.

7.2 Explainability Faithfulness

The Fuzzy-SHAP bridge is based on SHAP DeepExplainer, which approximates Shapley values with a finite background set, leading to approximation error in low-density areas [10]. The SHAP attributions to linguistic tokens mapping is carried out separately on each feature without considering fuzzy interactions between rules. This limitation could be overcome by a game-theoretic Fuzzy Shapley Value decomposition that takes

advantage of the explicit rule structure to give a more plausible account of multi-feature fuzzy rule activations.

7.3 Federated and Privacy-Preserving Extensions

The Fuzzy-SHAP bridge is based on SHAP DeepExplainer, which approximates Shapley values with a finite background set, leading to approximation error in low-density areas [10]. The SHAP attributions to linguistic tokens mapping is carried out separately on each feature without considering fuzzy interactions between rules. This limitation could be overcome by a game-theoretic Fuzzy Shapley Value decomposition that takes advantage of the explicit rule structure to give a more plausible account of multi-feature fuzzy rule activations.

7.4 Adversarial Robustness

The strength of UA-FNN against crafting adversarial transactions has never been formally tested in this study. The fuzzy membership structure can offer a potential smoothness of decision boundaries, but this suggestion must be strictly tested. Directions that can be taken in adversarial training include projected gradient descent adversarial training [30], and certified robustness via interval bound propagation through the fuzzy layers.

8 Conclusion

The paper has presented UA-FNN, an Uncertainty-Aware Fuzzy Neural Network to explainable risk scoring in real-time payment transactions. Using a combination of ANFIS-inspired fuzzification, an LSTM-based temporal consequent network, Monte Carlo Dropout uncertainty estimation, and the new Fuzzy-SHAP bridge, UA-FNN has been shown to meet simultaneously five requirements never met by a previous system: high discriminative performance, calibrating uncertainty quantification, linguistically interpretable explanations, fuzzy reasoning on transaction semantics, and sub-10 ms latency.

Empirical analysis on three benchmark datasets, an extensive ablation analysis on eight architectural variants, and statistical significance testing on DeLong and Wilcoxon tests show UA-FNN has state-of-the-art performance (AUC-ROC 0.971, 0.983, 0.977; ECE 0.031) within the latency limits of production payment processing. It is corroborated by the uncertainty routing model, which indicates that the routing of high-uncertainty transactions is an activity that has a 20 to 32 times higher probability of fraud than the default rate of the baseline, and that predictive uncertainty is a tractable risk management indicator, and has quantifiable operational value.

When the EU AI Act officially takes effect, covering high-risk AI systems in August 2026, financial institutions using automated payment scoring systems will have a tangible deadline on explainability and risk management compliance. The Fuzzy-SHAP explanations, calibrated uncertainty bands, and clear structure of rules of UA-FNN are specifically created to meet Article 86 (right to explanation), Article 12 (logging and documentation), and Article 9 (risk management system) requirements, making it a

technically sound and regulation-compliant architecture to next-generation production payment risk engines.

9 References

1. [1] IJSRA. (2025). Artificial intelligence and machine learning in fraud detection. International Journal of Scientific Research and Applications. https://ijsra.net
2. [2] He, L., Leslie, D., & Grant, J. (2024). An online fuzzy fraud detection framework for credit card transactions. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2024.123799
3. [3] Habibpour, M., Gharoun, H., Mehdipour, M., & Tajally, A. (2023). Uncertainty-aware credit card fraud detection using deep learning. Engineering Applications of Artificial Intelligence, 123, 106248.
4. [4] Jang, J.-S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics, 23(3), 665–685.
5. [5] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797.
6. [6] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.
7. [7] Cao, R., Wang, J., Mao, M., Liu, G., & Jiang, C. (2023). Feature-wise attention based boosting ensemble method for fraud detection. Engineering Applications of Artificial Intelligence, 126, 106975.
8. [8] Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: From anomaly detection to risk management. Financial Innovation, 9, 66.
9. [9] Duan, Y., et al. (2024). CaT-GNN: Enhancing credit card fraud detection via causal temporal graph neural networks. arXiv:2402.14708.
10. [10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
11. [11] Hjelkrem, L. O., & de Lange, P. E. (2023). Explaining deep learning models for credit scoring with SHAP. Journal of Risk and Financial Management, 16(4), 221.
12. [12] Global Risk Institute. (2023). Building explainable machine learning systems in finance: Fuzzy logic-based neural networks. GRI Technical Report.
13. [13] Thuy, A., & Benoit, D. F. (2023). Explainability through uncertainty: Trustworthy decision-making with neural networks. European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2023.09.009
14. [14] Ahmed, I. E., Mehdi, R., & Mohamed, E. A. (2023). The role of artificial intelligence in developing a banking risk index: An application of ANFIS. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10473-9
15. [15] Ayed, N., & Bougatef, K. (2024). Performance assessment of LR, ANN, FIS and ANFIS in predicting default probability. Computational Economics, 64, 1803–1835.
16. [16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. ACM SIGKDD, 1135–1144.
17. [17] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353.
18. [18] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. JAIR, 16, 321–357.

19. [19] Dwivedi, R., et al. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9), 1–33.
20. [20] European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (AI Act). Official Journal of the European Union.
21. [21] Cheng, D., et al. (2024). Enhancing attribute-driven credit card fraud detection with risk-aware graph representation. IEEE Transactions on Knowledge and Data Engineering.
22. [22] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. ICML 2016, 48, 1050–1059.
23. [23] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS 2017, 30.
24. [24] Angelopoulos, A. N., & Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning, 16(4), 1–140.
25. [25] Arrieta, A. B., et al. (2025). Model-agnostic explainable AI methods in finance: A systematic review. Artificial Intelligence Review. https://doi.org/10.1007/s10462-025-11215-9